**Supplementary Information**

Sections:
1. Analysis of dairy cattle data (FIG 3a).
2. Simulation based on the Drosophila melanogaster Genetic Reference Panel (FIG 3b).
3. Derivation of expected $R^2$ when discovery data are used both for SNP selection and prediction (BOX 2).
4. Analysis of height from the Framingham Heart Study (BOX 3)

***1. Analysis of dairy cattle data (FIG 3a)***

2,732 dairy bulls with ~509,096 genome wide SNP genotypes and with phenotype average milk yield of their daughters' milk production (all bulls had at least 50 daughters), a trait with "heritability" ~ 0.8. The bulls were split into a discovery sample (bulls born during or before 2003), n = 2,458, and a validation sample (bulls born after 2003 and related to bulls in the discovery sample) of n=274.

The analysis generating the blue line of FIG 3a was conducted in 3 steps. Step 1: A genome-wide association analysis was performed in the discovery sample. The model fitted to the data was $\mathbf{y} = \mathbf{1_n}'\mu + \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$; where **y** is the vector of daughter yield deviations, $\mathbf{1_n}$ is a vector of 1s of length n sires; **X** is a vector allocating SNP, *b* is the fixed effect of the SNP, **u** is the vector of polygenic breeding values, sampled from the distribution N(**0**, $\mathbf{A}\sigma^2$) where **A** is the average relationship matrix derived from pedigree (i.e. the dairy bulls are related including father-offspring, full- and half-sibling and more distant relatives) and **e** is the vector of random deviates. The analysis was carried out using ASReml[80]. SNPs with P values of $P<1x10^{-8}$, $P<1x10^{-6}$, $Px10^{-4}$ or $P<10^{-2}$ were taken to the next step (estimation of effect sizes).

Step 2: Genomic predictions were performed using GBLUP[36,81] to estimate genetic variances associated with the SNP sets identified in step 1. The following model was fitted to the data:

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Zg} + \mathbf{e}$$

where **y** is a vector of phenotypes (daughter yield deviations) of the discovery animals, $\mathbf{1_n}$ is a vector of 1s**,** μ is an overall mean, **Z** is a design matrix allocating records to breeding values, **g** is a vector of genomic breeding values and **e** is a vector of random normal deviates $V(e) \sim N(0,\sigma_e^2)$, where $\sigma_e^2$ is the error variance. The variance of breeding values was $V(\mathbf{g}) = \mathbf{G}\sigma_g^2$, where **G** is the genomic relationship matrix[26], using the SNP significant in the association analysis described above at the four different levels of significance and $\sigma_g^2$ is the genetic variance associated with the SNP set.

Step 3: breeding values (estimated genetic values which are hence the best estimate of the phenotypic values) for both discovery and validation individuals can be predicted as:

$$\hat{\mathbf{g}} = \left[\mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\frac{\sigma_e^2}{\sigma_g^2}\right]^{-1}\left[\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})\right]$$

Variance components were estimated with ASReml[80]. The proxy for accuracy of the genomic predictions used was $r(\hat{\mathbf{g}}, \mathbf{y})$ for the validation animals only, i.e., the correlation between true and estimated phenotype. The analysis generating the red line of FIG 3c was the same as above but the validation sample was included in step 1 (as well as step 3). The analysis generating the orange line of FIG 3c was the same as above but the validation sample was included in steps 1 and 2 (as well as step 3). The number of SNP detected in the association analysis at the different significant levels

was greater when the phenotypes and genotypes of the validation individuals were included, Table S1.

**Table S1. Number of SNP significant in GWAS for milk production**

| P-value threshold | Individuals used in association analysis | |
|---|---|---|
| | Discovery only | Discovery and validation |
| $1x10^{-8}$ | 631 | 688 |
| $1x10^{-6}$ | 948 | 1161 |
| $1x10^{-4}$ | 2820 | 3357 |
| $1x10^{-2}$ | 23337 | 25850 |

***2. Simulation based on the Drosophila melanogaster Genetic Reference Panel (FIG 2b)***

We downloaded the data reported in Mackay *et al*[54] from http://dgrp.gnets.ncsu.edu/. The data comprised162 inbred lines of Drosophila melanogaster each with ~4.7M markers. We filtered the marker data retaining only SNPs that were biallelic, autosomal, with minor allele frequency > 0.02 and with missingness < 0.1, leaving 1.96M autosomal SNPs. We excluded 8 lines because of missingness > 0.1, leaving 154 lines. Simulated phenotypes were generated from a normal distribution irrespective of the genotype data. We selected the top 10 associated SNPs by a multiple SNP association approach in GCTA[59] and predicted the phenotypes using these 10 SNPs to generate FIG 2b.

***3. Derivation of expected $R^2$ when discovery data are used both for SNP selection and prediction (BOX 2).***

We derive an approximation of the squared correlation between phenotype and predicted phenotype from SNP data in a sample size of $N$ unrelated individuals, when there is no correlation in the population between SNPs and phenotypes. Since discovery and validation sample are the same, $N = N_d = N_v$. The predictor is the sum of the product of estimated SNP effects and allele counts on $m$ out of $M$ selected SNPs. All SNPs are independent and their effect sizes are estimated in the same data by fitting them one at a time.

Our model is $y \sim N(0,1)$ and SNP effect are estimated as $y$ = mean + $b*x$. We can standardise $x$ to have a mean of 0 and a variance of 1, so that $b_i = R_i$. The best $m$ SNPs are selected on test statistic and the predictor is calculated as $\hat{y} = \sum_{i=1}^{m} R_i x_i$ . We are interested in $R^2 = R^2_{y,\hat{y}}$ , the amount of variance in $y$ spuriously "explained" by the predictor. Note that $\text{var}(R_i) = 1/N$. Under the null hypothesis the $i$-th ranked SNP out of $M$ tests has a p-value of $p_i = i/M$. Using $N*R^2 \sim \chi^2$ (with 1 df) the expected value of the proportion of variance explained by the i-th SNP, when fitted by itself, is

$E(R_i^2) = \chi^2_{[p_i]}/N$, with $\chi^2_{[p_i]}$ the $\chi^2$ value corresponding to a p-value of $p_i$.

$$R^2_{y,\hat{y}} = \frac{\text{cov}^2_{y,\hat{y}}}{\text{var}(\hat{y})}$$

$$\text{cov}_{y,\hat{y}} = \text{cov}(y, \sum R_i x_i) \approx \sum E(R_i^2)$$

An alternative expression for this sum can be derived from truncation normal theory, since the $R_i$ values for SNPs are selected from the lower and upper tail of the distribution from all $M$ estimates.

Since the variance of $R_i$ among all $M$ markers is $1/N$ and a proportion of $(m/2)/M$ is selected from each tail,

$\sum R_i^2 \approx$ m * var($R$|selection in one tail) = $(m/N) * (1 + i*t)$,

with $t$ and $i$ the truncation point and selection intensity for a proportion selected of $(m/2)/M$. This prediction agrees very well with simulations (not shown). For the case of $M = m$, the prediction is $m/N$, which is also the expected $R^2$ value from fitting m random covariates in a sample of size $N$ (assuming $m < N$) (Wishart 1931)[82].

$\text{var}(\hat{y}) = \text{var}[\Sigma\, R_i * x_i)]$

$= \Sigma[R_i^2 * \text{var}(x)] + \sum_i \sum_{j,i \neq j} (\text{cov}(R_i * x_i, R_j * x_j)$

$\approx [\Sigma\, E(R_i^2)] + \sum_i \sum_{j,i \neq j} (\text{cov}(R_i * x_i, R_j * x_j)$

For the second term, we use that $\text{cov}(R_i*x_i, R_j*x_j) = R_iR_j\ \text{cov}(x_i, x_j)$. $x_i$ and $x_j$ are independent except that both are correlated with $y$. Then $\text{cov}(x_i, x_j) = R_iR_j$ and so $\text{cov}(R_i*x_i, R_j*x_j) = (R_iR_j)^2$. Hence,

$\Sigma\Sigma(\text{cov}(R_i * x_i, R_j * x_j) = \Sigma\Sigma(R_iR_j)^2 = [\Sigma(R_i^2)]^2 - \Sigma(R_i^4)$

The second term is generally much smaller than the first one, so we can approximate $\Sigma\Sigma(\text{cov}(R_i * x_i, R_j *) \approx [\Sigma(R_i^2)]^2 \approx [(m/N) * (1 + i*t)]^2$, and

$\text{var}(\hat{y}) \approx (m/N) * (1 + i*t) + [(m/N) * (1 + i*t)]^2$

$= (m/N) * (1 + i*t)[1 + m/N) * (1 + i*t)]$

Putting the analytical results together,

$R^2_{y,\hat{y}} \approx [m(1 + i*t)] / [N + m(1 + i*t)]$

This result was validated by simulations for N = 500, 1000 and 5000, M = 1000, 10000 and 5000, and $m$ = 10, 100 and 1000. The predicted values were within 0.005 of the average value from 100 replicate simulations.

### *4. Analysis of height from Framingham Heart Study (BOX 3)*

**Table S2.** Impact of relatedness and stratification on FHS polygenic prediction. For each relatedness threshold, we list prediction $R^2$ for height and ancestry-adjusted height (adjusted for 10 eigenvectors). In the run with no close relatives, no relatedness threshold was imposed at the training stage but prediction $R^2$ was computed for a subset of 1,880 individuals that had no close relatives in the entire data set, according to pedigree information.

| | # training samples (10-fold cross-valid) | # validation samples (10-fold cross-valid) | prediction $r^2$ for height | prediction $r^2$ for anc-adjusted height |
|---|---|---|---|---|
| $x$=0.05 threshold | 1,797 | 1,997 | 0.059 | 0.000 |
| $x$=0.20 threshold | 2,228 | 2,475 | 0.062 | 0.012 |
| $x$=0.40 threshold | 2,809 | 3,121 | 0.075 | 0.064 |
| No known close relatives | 6,691 | 1,880 | 0.129 | 0.185 |
| All samples | 6,691 | 7,434 | 0.144 | 0.263 |

$x$ = maximum relationship threshold based on genome-wide genotypes.

The results in the table demonstrate a subtle issue not discussed in the main paper. Firstly, although the decreasing $R^2$ as a function of relatedness cut-off (0.129, 0.075, 0.062, 0.059) could in theory be due to reduced sample size, the much sharper decreases in the ancestry-adjusted analysis in the rightmost column (0.185, 0.064, 0.012, 0.000) imply that this is due to relatedness and not reduced sample size. However, note that in the bottom 2 rows, ancestry adjustment actually increases prediction $R^2$. This is because when the prediction $R^2$ is high, not adjusting for ancestry in an analysis that fit markers individually will overweight ancestry in the prediction, pulling the prediction $R^2$ down.